VLOGGER: Multimodal Diffusion for Embodied Avatar Synthesis

Enric Corona Andrei Zanfir Eduard Gabriel Bazavan Nikos Kolotouros
Thiemo Alldieck Cristian Sminchisescu

Google Research

Figure 1. We present VLOGGER, a novel framework to synthesize humans from audio. Given a single input image like the ones shown on the first column, and a sample audio input, our method generates photorealistic and temporally coherent videos of the person talking and vividly moving. As can be seen on the synthesized images in the right columns, we generate head motion, gaze, blinking, lip movement and unlike previous methods, upper-body and hand gestures, taking audio-driven synthesis one step further.

Abstract

*We propose VLOGGER, a method for text and audio-driven talking human video generation from a single input image of a person, which builds on the success of recent generative diffusion models. Our method consists of 1) a stochastic human-to-3d-motion diffusion model, and 2) a novel diffusion-based architecture that augments text-to-image models with both temporal and spatial controls. This approach enables the generation of high quality videos of variable length, which are easily controllable through high-level representations of human faces and bodies. In contrast to previous work, our method does not require training for each person, does not rely on face detection and cropping, generates the complete image (not just the face or the lips), and considers a broad spectrum of scenarios (*e.g. *visible torso or diverse subject identities) that are critical to correctly synthesize humans who communicate. We evaluate VLOGGER on three different benchmarks, and show that the proposed model surpasses other state-of-the-art methods in image quality, identity preservation and temporal consistency. We collect a new and diverse dataset MENTOR, one order of magnitude larger than previous ones (2,200 hours and 800,000 identities, and a test set of 120 hours and 4,000 identities) on which we train and ablate our main technical contributions. We report the performance of VLOGGER with respect to multiple diversity metrics, showing that our architectural choices benefit training a fair and unbiased model at scale.*

1. Introduction

We present VLOGGER, a method to automatically generate a video of a talking person, based on text or audio, and given only a single image of that person. Industries like content

creation, entertainment, or gaming all have high demand for human synthesis. Yet, the creation of highly realistic videos of humans is still complex and ripe with artifacts, thus requiring significant manual intervention for realistic results. Full automation, however, would not only ease creative processes, but also enable entirely new use cases, such as enhanced online communication, education, or personalized virtual assistants, to name a few. The latter is especially relevant, given the recent success of chat agents [47, 55]. Research has shown that such solutions are not perceived as natural enough to develop empathy [104] and several authors [40] argue that anthropomorphism and behavioral realism (*e.g*. eye gaze, facial expressions, natural whole-body movements *etc*.) are critical in creating a social presence and in eliciting empathetic responses from the user. Such features would result in the wide adoption of agents [50], in areas like customer service [1, 57], telemedicine [66], education [65], or human-robot interaction [62]. It is precisely automation and behavioral realism that what we aim for in this work: VLOGGER is a multi-modal interface to an *embodied conversational agent* [76], equipped with an audio and animated visual representation, featuring complex facial expressions and increasing level of body motion, designed to support natural conversations with a human user. VLOGGER can be used as a stand-alone solution for presentations, education, narration, low-bandwidth online communication, and as an interface for text-only HCI [4, 101].

Photorealistic human synthesis is a complex task due to challenges like data acquisition, enacting facial expressions in a natural way, expression to audio synchronization, occlusion, or representing full-body movements — especially given a single input image. Many attempts focused exclusively on lip sync [58, 77, 84], in which they edit the mouth region of a driving video. Recently, some works [94, 96] relied on the extensive advances in face reenactment [9, 20, 35, 54, 71, 88, 97] to generate talking head videos from a single image, by predicting face motion from audio. In doing so, they rely on flow-based audio to expression networks to animate the original input image. However, we find that such approaches cannot completely disentangle the expression and facial attributes of the input when generating a new video. Temporal consistency is usually achieved with a per-frame image generation network by relying on a smooth guiding motion from face keypoints. However, this might cause blurriness and does not ensure temporal coherency in areas of the head more distant from the face. Consequently, most methods require detecting and cropping the head when a significant part of the body is visible. In this paper, we argue that communication is more than "just" audio combined with lip and face motion – humans communicate using their body via gestures, gaze, blinking, or pose. MODA [43] recently started exploring the animation of both face and body, however in limited scenarios, and without generalizing to new identities. In contrast, we aim for a *general, person agnostic synthesis solution*.

Text Control	Audio Control	Face Control	Body Control	Stochastic	Photorealistic	Generalizes to new subjects	Resolution	
✗	✗	✓	✗	✗	✓	✓	-	Face Reenactment [71, 81]
✗	✓	✓	✗	✓	✗	✓	-	Audio-to-Motion [19, 70]
✗	✓	✗	✗	✗	✓	✓	-	Lip Sync [24, 58]
✗	✓	✓	✗	✗	✓	✓	256	SadTalker [96]
✗	✓	✓	✗	✗	✓	✓	256	Styletalk [46]
✓	✓	✓	✓	✓	✓	✓	512	**VLOGGER (Ours)**

Table 1. **Key properties of VLOGGER compared to related work**. Face Reenactment [9, 20, 35, 54, 71, 88, 97] generally does not consider driving audio or text. Work on audio-to-motion [15, 19, 61, 67, 70, 86, 91] shares components by encoding audio into 3d face motion, however lack realism. Lip sync [24, 58] consider input videos of different subjects, but only model mouth motion. Given their generalisation capacity, SadTalker [96] and Styletalk [46] are the closest to us, but require cropped images of faces, lack body control, and are lower resolution.

Towards that goal, we propose a two-step approach where first a generative diffusion-based network predicts head and body motion according to an input audio signal. This stochastic approach is necessary to model the nuanced (one-to-many) mapping between speech and pose, gaze, and expression. Second, we propose and ablate a novel architecture based on recent image diffusion models, which provide controls in the temporal and spatial domains. By additionally relying on generative human priors, acquired during pre-training, we show how this combined architecture improves the capacity of image diffusion models, which often struggle to generate consistent human images (*e.g*. eyes). Finally, we produce a base model followed by a super-resolution diffusion model to obtain high quality videos. We emphasize the multi-modality of our approach, which can use text and audio controls, and a single image, to generate video of new content. In doing so we connect image generative models and text-to-speech methodologies towards conversational agents that generalise well. For robustness and generalisation, we collect a large-scale dataset featuring a much larger diversity than previously available data, in terms of skin tone, body pose, viewpoint, speech and body visibility. We show that our model, when trained on this data, outperforms previous work across different diversity metrics, and obtains state-of-the-art image quality and diversity results on the previous HDTF [99] and TalkingHead-1KH [81] datasets. Moreover, our method considers a larger range of scenarios than previous works, by generating high resolution video of head and upper-body motion.

2. Related Work

Audio-Driven Talking Face Generation. There has been a significant amount of work in talking face generation, which can be categorized according to the driving inputs, intermediate representations and output formats. We provide an overview and comparison against our work in Tab. 1. There exists a body of work in animation of 3D morphable faces [15, 19, 61, 67, 70, 86] or full body [91] models based on audio segments. These efforts can generate diverse 3d talking heads in the form of temporally coherent pose and shape parameters of various statistical head or body models [6, 7, 41, 56, 87]. We consider a similar network to guide the generated motion but, in this paper, we instead aim to generate photorealistic talking humans with diversity in expression and head motion, that are coherent with an input image of a target subject. We consider challenges such as temporal consistency, diversity in cloth style, hair, gaze, and detail in output videos.

In the image domain, incipient works have focused on the task of mouth editing [10, 13, 36, 58, 75, 99], such as only predicting the lip motion, synchronized with the input audio. Follow up works added extended features such as head motion, gaze and blinking, [37, 45, 60, 69, 98, 103] using intermediate 2d, 3d landmarks or flow based representations. To increase the level of photorealism, a large number of works have extensively used discriminators as part of the losses [8, 9, 18, 59, 82, 93], and some recent methods proposed the use of diffusion models [67, 68, 94]. However, it is hard to ensure proper disentanglement between body, head motions, blinking, gaze and facial expressions when operating in the latent space of GANs [21, 38] or generic diffusion models. Our method does not need to employ custom perceptual, gaze, identity preserving or lip syncing losses.

Body motion and gestures have not been considered because of the lack of data and the difficulty of generating coherent video. We collect a large-scale dataset and propose a complete pipeline that can take as input guiding text or audio. It can generate coherent face and upper-body motion, with a variety of facial expressions, head motion, gaze, eye blinking and accurate lip movement synced with the input audio. Moreover, we show that our method is more expressive and robust across different diversity axis.

Face Reenactment. Video-based talking face generation aims to transfer the motion of a source video to a target person, and has been widely explored in the past [9, 26, 34, 35, 54, 71, 83, 88, 97, 100, 102]. Most methods rely on an intermediate representation, such as sparse or dense landmarks, semantic masks, 3d dense representations or warped features. In the 3d domain, several works have taken advantage of NeRF [5, 48] based solutions [25, 42, 89, 90]. However, this requires a significant amount of frames of a target person talking, for retraining and animating them. This task is closely related to ours, and some previous works adapt these intermediate representations when considering audio as input. In our case, however, we aim to move forward from face-only videos and consider more diverse input samples, *e.g.* containing body and hair motion.

Video Generation. Also related to our work is the topic of video generation. This is a task that has been widely explored in the community, thus we only focus on the most related directions. With the success of text-to-image diffusion models [17], many works have also explored their extension to the video domain [27, 32, 39, 74, 85] but are limited in number of seconds or resolution. Moreover, most previous works do not explicitly tackle humans despite the amount of data available. In our case, we extend current state-of-the-art image diffusion models to the temporal domain by adding spatio-temporal controls and propose an iterative outpainting procedure to generate videos of variable length.

3. Method

Our goal is to generate a photorealistic video $\mathbf{V}$ of variable length depicting a target human talking, including head and gestures. Our framework, which we call VLOGGER, is illustrated in Fig. 2. VLOGGER is a two-stage pipeline based on stochastic diffusion models to model the one-to-many mapping from speech to video. The first network takes as input an audio waveform $\boldsymbol{a} \in \mathbb{R}^{NS}$ at sample rate S to generate intermediate body motion controls $\mathbf{C}$, which are responsible for gaze, facial expressions and pose over the target video length. The second network is a temporal image-to-image translation model that extends large image diffusion models, taking the predicted body controls to generate the corresponding frames. To condition the process to a particular identity, the network also takes a reference image of a person. We train VLOGGER on our newly introduced, large-scale MENTOR dataset (see §3.3). We describe both networks next.

3.1. Audio-Driven Motion Generation

Architecture. The first network of our pipeline M is designed to predict the driving motion, and takes input controls from multiple domains, like audio or text. For text, we employ a text-to-speech model to convert inputs to waveforms [72], and represent the resulting audio as standard Mel-Spectrograms. M is based on a transformer architecture [73] with four multi-head attention layers on the temporal dimension. We include positional encoding on the number of frames and diffusion step, and an embedding MLP for the input audio and the diffusion time step. At each frame, we use a causal mask so that the model can only attend to previous frames. The model is trained using variable length videos to enable generation of very

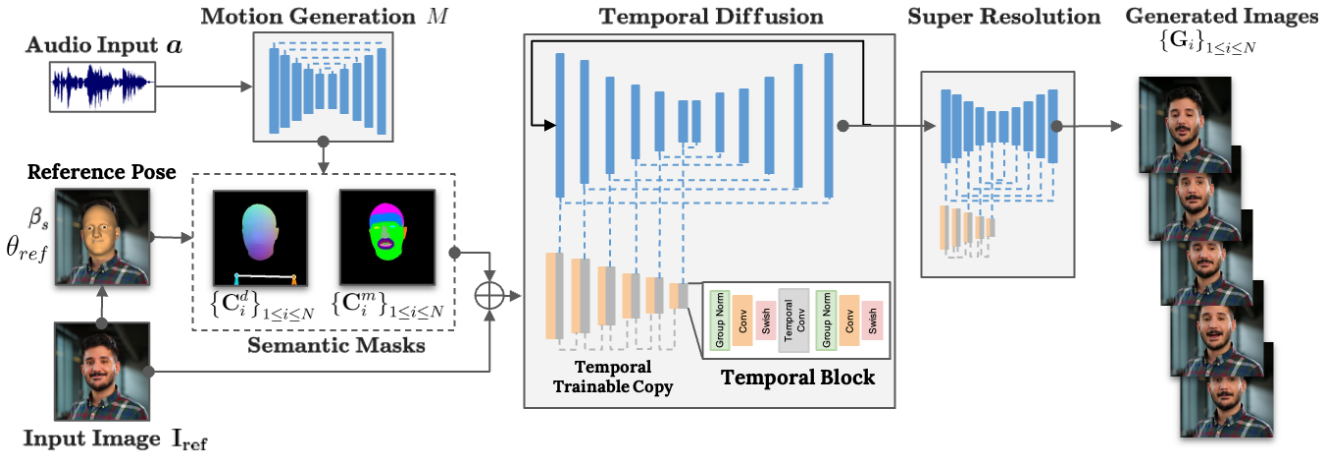


Figure 2. **High-level overview of VLOGGER.** For an input image $\mathbf{I}_{\mathbf{ref}}$, we fit a statistical head model SPHEAR [6] and recover the shape parameters β_s of the target identity. We pass the audio input Mel-Spectrogram $\mathbf{a}$ to our motion network to generate a sequence of facial expressions $\{\theta_i^e\}_{1\leq i\leq N}$, head poses $\{\theta_i^h\}_{1\leq i\leq N}$ and body poses $\{\theta_i^b\}_{1\leq i\leq N}$. Using the SPHEAR parameters, we render a sequence of dense facial expression image controls $\{\mathbf{C}_i^d\}_{1\leq i\leq N}$, as well as semantic masks $\{\mathbf{C}_i^m\}_{1\leq i\leq N}$ for relevant facial regions. We pass all generated control images and the target image $\mathbf{I}_{\mathbf{ref}}$ as inputs to a temporal diffusion model and a super resolution module, which are trained to generate a sequence of photorealistic reenactments $\{\mathbf{G}_i\}_{1\leq i\leq N}$ of the target identity.

long sequences, as *e.g.* in the TalkingHead-1KH Dataset [81] (see §4). We rely on the estimated parameters of a statistical 3d head model, SPHEAR [6], together with 3d body pose [22], to produce intermediate control representations (dense or sparse, in image space) for the synthesized video. The motion generation is then tasked with predicting $M(\boldsymbol{a}_i) = \{\theta_i^e, \Delta\theta_i^h, \Delta\theta_i^b\}$. More precisely, the model predicts facial expression parameters $\theta_i^e \in \mathbb{R}^{64\times 1}$, residuals over the head pose, *i.e.* $\Delta\theta_i^h \in \mathbb{R}^{9\times 1}$, and residuals over the body pose, *i.e.* $\Delta\theta_i^b \in \mathbb{R}^{75\times 2}$. By predicting displacements, *i.e.* $\Delta\theta_i^h$ and $\Delta\theta_i^b$, we enable the model to take an input image with reference poses θ_{ref}^h and θ_{ref}^b for the target subject, and animate the person relatively with $\theta_i^h = \theta_{\text{ref}}^h + \Delta\theta_i^h$ and $\theta_i^b = \theta_{\text{ref}}^b + \Delta\theta_i^b$, for frames $1 \leq i \leq N$. The predicted model parameters, *i.e.* θ_i^e and θ_i^h, are used to pose the speaker's 3d head and body representation. The SPHEAR model also requires a head shape code, $\boldsymbol{\beta}_s \in \mathbb{R}^{64\times 1}$, which encodes the geometric identity of the speaker. During training, we use the (estimated) ground-truth value, while at test time we use the value retrieved from fitting the parametric model on the input image. In order to leverage the 2d/3d predictions with CNN-based architectures, we rasterize the template vertex positions of the posed heads as dense representations, and body as sparse joint representations, to obtain dense masks $\{\mathbf{C}_i^d\}_{1\leq i\leq N} \in \mathbb{R}^{H\times W\times 3}$. We also rasterize the semantic regions of the posed heads, $\{\mathbf{C}_i^m\}_{1\leq i\leq N} \in \{0,1\}^{H\times W\times N_c}$, where N_c is the number of semantic classes. The rasterization process assumes a full-perspective camera, with a diagonal field-of-view inferred from either the training video, or the reference image. For illustrations, please see Fig. 2. The generated dense and semantic masks are used as conditional signals for our temporal image diffusion model, which we describe in more detail in the next section.

Loss functions. This model follows a diffusion framework which progressively adds Gaussian noise $\epsilon \sim \mathcal{N}(0,1)$ to the ground-truth samples $x_0 = \{\{\theta_i^e, \Delta\theta_i^h, \Delta\theta_i^b\}\}_{1\leq i\leq N}$, with a conditional audio input $\boldsymbol{a}$. The goal is to model the motion distribution of real heads and bodies, $x_0 \sim q(x_0|\boldsymbol{a})$, by training a denoising network ϵ_ϕ that predicts the added noise from the noisy input x_t, where t is an arbitrary diffusion step. In our case, we obtained better performance by directly predicting the ground-truth distribution:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0,t,\boldsymbol{a},\epsilon\sim\mathcal{N}(0,1)}\left[\|x_0 - \epsilon_\phi(x_t, t, \boldsymbol{a})\|_2^2\right] \quad (1)$$

We also include an additional temporal loss to obtain coherent results by penalizing prediction difference at consecutive frames, $\mathcal{L}_{\text{temp}} = \|\epsilon_\phi(x_t,t,\boldsymbol{a})_{i+1} - \epsilon_\phi(x_t,t,\boldsymbol{a})_i\|_2^2$, for any given frame $i \in N$, and train the full model using a linear combination of both losses, *i.e.* $\mathcal{L}_{\text{diff}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}$. In practice, we use different temporal loss weights for each part of the face to ensure smoother motion for the head while allowing larger dynamism for facial expressions.

3.2. Generating Photorealistic Talking Humans

Architecture. Our next goal is to animate an input image $\mathbf{I}_{\mathbf{ref}}$ of a person, such that it follows the previously predicted head motion, which is represented with semantic, sparse and dense masks $\mathbf{C}$. Based on these image-based controls, we propose a temporally-aware extension of state-of-the-art diffusion models, extending a version of Imagen [64] trained on internal data. Inspired by ControlNet [95], we freeze the initial trained model and make a zero-initialized trainable copy of its encoding layers, which take the input temporal controls $\mathbf{C}$. We interleave 1d convolutional layers in the temporal domain, after the first layer

Metrics in the final video	FID [29] ↓	LME [mm] ↓	Jitter [mm/s^3] ↓
	Motion Generation		
Not predicting Δ	52.27	4.22	6.56
No temporal loss	16.56	3.18	4.64
No classifier-free guidance	16.54	3.32	**3.49**
	Temporal Diffusion Model		
No body landmarks as input	16.95	3.10	4.45
No temporal outpainting	15.88	3.25	3.70
25% outpainting overlap	15.90	3.23	3.61
Full model	**15.36**	**3.06**	3.58

Table 2. **Ablation study of the main design choices in VLOGGER** evaluated on the MENTOR Dataset, where we report the most representative metrics to validate image quality through the FID [29] score, expressiveness and lip sync quality via landmark error (LME), and temporal consistency based on face vertex jitter. In the first part, we show that the temporal loss and use of classifier-free guidance lead to the best performance in image quality and LME (Full model in last row). For the temporal diffusion model, we show our final model benefits from taking body landmarks, and validate the effectiveness of the proposed temporal outpainting, with the full model (at 50% of outpainting overlap) obtaining the best temporal consistency. We noticed the model plateaus with more overlap.

of each downsampling block and before the second GroupNorm activation, as shown in Fig. 2. The network is trained by taking N consecutive frames and controls, and tasked to generate short clips of the reference person animated according to the input controls.

Training. We train our method on the MENTOR dataset, which consists of full-length videos of unique human subjects. Because, during training, the network takes a sequence of consecutive frames and an arbitrary reference image $\mathbf{I}_{\mathbf{ref}}$ of the person, we theoretically can assign any video frame as the reference. In practice, we sample the reference to be farther away (temporally) from the target clip, as closer examples trivialize the training and provide less generalization potential. We train all image models with $\texttt{lr} = 5e-5$, with the base and super-resolution network being trained for $280k$ steps with $\texttt{bsize} = 128$. More details about the training procedure are provided in Supp. Mat.

Loss functions. Similar to the previous section and the loss described in Eq. (1), we follow a diffusion process in which we add noise ϵ^I to the ground-truth images $\mathbf{I}$. We base our work on a version of Imagen [64] trained on internal data sources, which predicts the added noise ϵ^I:

$$\mathcal{L}^I_{\text{diff}} = \mathbb{E}_{x^I_0,t,\mathbf{C},\epsilon^I\sim\mathcal{N}(0,1)}\left[\left\|\epsilon^I - \epsilon^I_\phi(x^I_t, t, \mathbf{C})\right\|^2_2\right] \quad (2)$$

Super Resolution. While the previous approach is resolution independent, we generate base videos at 128×128 resolution, and use a cascaded diffusion approach to extend the temporal conditioning in two super-resolution variants for higher quality video at 256×256 or 512×512. The generated images are denoted as $\{\mathbf{G}_i\}_{1\le i\le N}$. High resolution examples are shown in Fig. 1 and Fig. 4.

Temporal outpainting during inference. The proposed temporal diffusion model is trained to generate only a fixed number of frames N, so it is not obvious how to extend it to variable length video. Most previous diffusion-based video generation methods are limited to short clips [33, 39, 85] or rely on smoothly generated intermediate token representations [74], but without guarantees of smooth changes in the pixel domain. Here, we explore the idea of temporal outpainting: we first generate N frames, and then we iteratively outpaint $N' < N$ frames based on the previous $N - N'$. The amount of overlap between two consecutive clips, *i.e.* $N - N'$ is chosen as a trade-off between quality and running time. We use DDPM to generate each video clip, and show that such approach can scale to thousands of frames. For details, see the ablation in Tab. 2, where we validate the main design choices and show that our final network can generate realistic and temporally coherent higher resolution video of humans.

	PSNR ↑	SSIM ↑	LPIPS ↓	L1 ↓
Ours	22.36	0.769	0.0782	.0426
	Results across visibility (Ours)			
Tight Face	24.18	0.8084	0.065	.033
Head & Torso	23.18	0.778	0.070	.0398
Torso & Hands	20.83	0.759	0.0839	.0461

Table 3. **Full-body reenactment results** in the MENTOR Dataset. We report image similarity metrics on average, and computed separately for images with only visible face, torso or hands.

3.3. MENTOR Dataset

We select a corpus of online videos that contain a single speaker, mostly facing the camera, from the torso up, communicating mostly in English. We split the videos in 240 frames at 24 fps (10 seconds clips) for easy processing. The audio is at 16 kHz synchronized with ground-truth timestamps to the video signal.

Fitting. To obtain accurate head estimations that are temporally smooth, we leverage the recently introduced statistical head model SPHEAR [6], which we fit to the videos in a two-stage pipeline. We first rely on an in-house facial landmark detector that provides a sparse semantic signal per-frame. Next, we use a trained landmark lifter that goes from 2d landmarks to initial head parameters (similar to [22, 23]). We use a non-linear quasi-Newton optimization (L-BFGS) framework, with temporal constraints and a fixed shape parameter per clip, to get the final model parameters. The losses include landmark re-projection error, latent prior regularization and temporal regularization. We adopt a full-perspective camera, with the diagonal field-of-view as an optimization variable.

Preprocessing. The videos in MENTOR also contain a significant part of the body, thus we estimate 3d body joints and hands. We filter out videos where the background changes meaningfully, the face or body have been only partially detected or their estimations is *jittery*, where hands are completely undetected (*e.g.* in cases of humans grasping and manipulating objects), or the audio is of low quality.

		HDTF Dataset [99]								
		Photorealism			Lip Sync		Diversity	Identity Preserv.		Temp. Consist.
	Is generative?	FID [29] ↓	CPBD [52] ↑	NIQE [49] ↓	LSE-D [12] ↓	LME [mm] ↓	Expression ↑	Head Err. ↓	ArcFace [16] ↓	Jitter [mm/s^3] ↓
Groundtruth Video	-	0.0	0.562	6.31	7.79	0.0	0.401	0.0	0.0	5.19
MakeItTalk [106]	✗	22.63	0.428	6.65	8.30	3.26	0.364	0.911	0.828	6.21
Audio2Head [79]	✗	19.58	0.512	**6.41**	**7.55**	3.08	**0.415**	0.896	1.92	6.15
Wang *et al.* [80]	✗	21.23	0.428	7.71	8.04	4.48	0.365	1.37	2.52	6.46
SadTalker [96]	Head Pose	**19.44**	**0.520**	6.48	**7.73**	**3.01**	0.287	**0.880**	0.874	5.51
StyleTalk [46]	✗	34.16	0.472	6.47	7.87	3.79	**0.416**	1.14	**0.692**	**4.34**
Ours	✓	**18.98**	**0.621**	**5.92**	8.10	**3.05**	0.397	**0.877**	**0.759**	**5.05**
Ours (Best of 3)	✓	-	0.628	5.64	7.43	2.95	0.425	0.829	0.706	4.75
Ours (Best of 5)	✓	-	0.631	5.53	7.22	2.91	0.436	0.814	0.687	4.67
Ours (Best of 8)	✓	-	0.634	5.44	7.04	2.84	0.448	0.800	0.677	4.58
		TalkingHead-1KH Dataset [81]								
		Photorealism			Lip Sync		Diversity	Identity Preserv.		Temp. Consist.
	Is generative?	FID [29] ↓	CPBD [52] ↑	NIQE [49] ↓	LSE-D [12] ↓	LME [mm] ↓	Expression ↑	Head Err. ↓	ArcFace [16]↓	Jitter [mm/s^3] ↓
Groundtruth Video	-	0.0	0.512	7.27	8.70	0.0	0.452	0.0	0.0	3.91
MakeItTalk [106]	✗	34.84	**0.493**	7.86	10.48	3.50	0.382	**1.20**	0.909	4.69
Audio2Head [79]	✗	46.49	0.475	7.55	**9.38**	4.33	**0.494**	1.47	2.01	4.66
Wang *et al.* [80]	✗	34.52	0.440	8.61	10.18	3.49	0.338	1.48	2.93	4.70
SadTalker [96]	Head Pose	**31.45**	0.482	**7.46**	**8.17**	**3.10**	0.347	1.21	0.961	4.26
StyleTalk [46]	✗	38.98	0.468	7.96	9.46	3.44	0.421	1.29	**0.663**	**3.19**
Ours	✓	**28.94**	**0.575**	**6.91**	9.40	**3.33**	**0.436**	**1.05**	**0.881**	**4.16**
Ours (Best of 3)	✓	-	0.582	6.33	8.969	3.07	0.448	1.03	0.853	3.68
Ours (Best of 5)	✓	-	0.585	6.21	8.93	2.96	0.455	1.01	0.833	3.57
Ours (Best of 8)	✓	-	0.589	6.08	8.90	2.94	0.469	0.99	0.813	3.56

Table 4. **Quantitative evaluation on the HDTF and TalkingHead-1KH Datasets.** We measure the capacity of our model to generate realistic talking heads in multiple metrics. VLOGGER achieves the highest visual quality with highest identity preservation summarized in several metrics, while obtaining expression diversity and temporal consistency close to the groundtruth videos. Regarding lip sync quality, all methods obtain comparable scores. To demonstrate the diversity generated by VLOGGER, we also report the improvement in performance when generating 3, 5 or 8 videos (Except for FID which measures a similarity within an image distribution). Results are consistent for all metrics on both datasets and the best two scores are marked in **bold**.

This process resulted in a training set of more than 8M seconds (2.2K hours) and 800K identities, and a test set of 120 hours and ∼4K identities, making it the largest available dataset to date in terms of identities and length, at higher resolution. Moreover, the MENTOR dataset contains a wide diversity of subjects (*e.g.* skin tone, age), viewpoints or body visibility. Statistics and a broader comparison to currently available datasets are provided in Supp. Mat. We aim to release the curated video ids, face fits and estimated body pose to the broader research community.

4. Experiments

Data and Training. We train VLOGGER on the proposed MENTOR dataset as described in Sec. 3.3, at a base resolution of 128 × 128 and cascade resolutions at 256 × 256 and 512 × 512. Evaluation is performed on the test sets of the MENTOR, HDTF [99] and TalkingHead-1KH [81] datasets. We also ablate the performance of our method in different scenarios on the MENTOR dataset and report its performance against baselines across several diversity metrics, such as age, perceived gender, or skin tone.

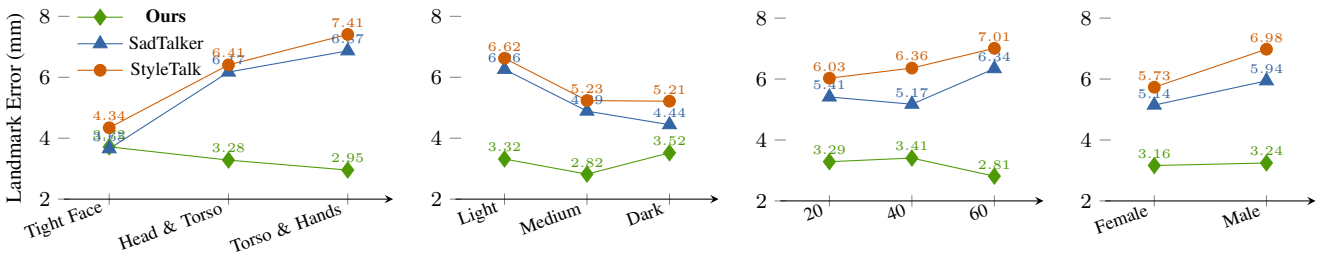
Baselines. We compare against several state-of-the-art methods, i.e. [46, 79, 80, 96, 106]. Note that, unlike our method, all baselines require cropping the face region, as they can detect and animate only the head.

Metrics. There is no single metric to evaluate the quality of generated conversational videos and, like previous work, we rely on a combination of them in order to evaluate image quality, lip sync, temporal consistency, and identity preservation. We first report the FID score [29] to measure the distance between ground-truth and generated distributions, the Cumulative Probability of Blur Detection (CPBD) [51, 52] and Natural Image Quality Evaluator (NIQE) [49] to validate the quality of generated images. We next fit the SPHEAR head model [6] to both ground-truth and generated videos, and report the difference in mouth vertex position (LME) to measure the lip sync quality. We also report the LSE-D [12] score. Similarly, we report the jitter (or *jerk*) error following [92] to measure the temporal smoothness in generated videos. We also provide the standard deviation of the parameters of the fitted SPHEAR head model to assess diversity in terms of expression and gaze, given that speech-to-video is not always a one-to-one mapping and it is important to generate a distribution of realistic videos.

4.1. Ablation Study

We first ablate our main design choices extensively in Tab. 2, by running our final model and by evaluating the most representative metrics in the generated videos. Each row evaluates the final model by only changing one feature (*e.g.* not using a temporal loss when training the motion predictor). We discuss the results next.

Motion generation. In the upper-part of Tab. 2 we show the drop in temporal consistency when not using temporal loss or not predicting Δ. By having the network predict



(a) Visible Body Parts **(b) Skin Tone** **(c) Age** **(d) Gender**

Figure 3. Our model and closest competitors across **different perceived attributes**, such as skin tone, gender and age, on the test set of MENTOR dataset. Our model leverages priors from large pre-trained diffusion models and our proposed large-scale dataset. Thus, in contrast to other methods, it manages to perform consistently across all categories, showing little to no bias. We also show in (a) that our model is capable of animating humans in images at a wide range of viewpoints, instead of cropping tight bounding boxes around the face.

a residual over the body motion, we gain in smoothness and stability, which results in overall higher image quality. Finally, we also show the positive use of classifier-free guidance (discussed more in depth in Supp. Mat.) regarding LME and FID [29] scores.

Video Generation. We also provide several examples to validate our current temporal video generation model. We show the effectiveness of the proposed outpainting procedure, which not only supports variable-length video generation, but also ensures smoothness and low jitter. Our final model has an overlap of 50% between generated and given frames, and plateaus at larger values, but obtains a noticeable improvement with respect to a smaller overlap (25%), or no outpainting. The model also performs better with body pose control.

Reenactment. We also report our performance in Tab. 3 on the MENTOR dataset for full-body reenactment in videos with tight-cropped faces, visible torso, or visible hands.

4.2. Quantitative Results

Talking Head Generation. Tab. 4 summarizes the performance of VLOGGER against previous state-of-the-art methods on the task of audio-driven video generation. We report results on the HDTF Dataset [99], a large scale dataset, but with not that many subjects (300) subjects and somewhat limited viewpoint variability, and on the TalkingHead-1KH Dataset [81]. Talking head generation is a challenging task with several desirable properties, assessed by different metrics. Noticeably, there is a trade-off between image quality, diversity and identity preservation. VLOGGER comes close to the amount of expression diversity present in real videos while achieving the highest image quality and identity preservation, with second lowest motion jitter after StyleTalk [46], which introduces very little face motion (see Fig. 4). The temporal consistency validates the contribution of our temporal layer and the outpainting procedure, while still leveraging the high-quality image generation capabilities of state-of-the-art diffusion models. All methods obtain comparable Lip Sync scores, and results are consistent for all metrics on both datasets evaluated. We also evaluate our method with different number of samples produced (3, 5 or 8) by selecting the best performing video per subject, leading to significantly improved performance with growing number of samples. These support the generative properties and diversity of VLOGGER.

In Fig. 3, we showcase our fairness and generalization capabilities (in part due to the scale and diversity of our training set), by running comparisons to other methods across several perceived attributes. Previous works exhibit a clear performance degradation for different classes (*e.g.* light vs dark skin, young vs old, *etc.*), and do not generalize to videos with visible torsos or hands. In contrast, VLOGGER exhibits fairly low bias on all the evaluated axes. We hope that the release of MENTOR will enable other researchers to address critical fairness issues in their work, and further advance the state-of-the-art.

4.3. Qualitative Results

We show qualitative results in Fig. 4 against the most recent and high-performing baselines on images in-the-wild. While most previous work rely on feature warping, this makes it difficult to generate parts occluded in the reference image (*e.g.* if the teeth were obscuring the mouth interior, they will persist across the generated video). In contrast, our model is able to generate more diverse expressions and correctly inpaint occluded regions of moving heads.

Sample diversity. Since VLOGGER is stochastic, we can generate multiple motions and videos given the same input audio/text, as illustrated in Fig. 5. From the first row, it can be seen that while the background is almost static, the face, hair, gaze and body motion feature an increasing amount of change as the video temporally unfolds. Even for the stochasticity shown in this example, all generated videos are of good quality (second row).

Personalization. Personalization in the context of diffusion models has been extensively explored recently for subject-driven generation [63]. In our case, VLOGGER only takes a monocular input image as source for synthesis, which may lead to a lack of visual detail or artifacts. While our model can produce a plausible synthesis, it has no access to occluded parts and the resulting video may not be veridical at a fine grain analysis of that person, if data

Figure 4. **Qualitative results and comparison against baselines**, showing the input image for each example (left) and frames from video synthesis. Most previous work rely on warping, which makes it difficult to modify attributes that are visible in the input image, *e.g.* mouth when teeth are visible (first and third rows). Similarly, most methods maintain the expression along the whole sequence or minimally modify the subject expression. Most noticeably, while previous work requires cropping the head [46, 80, 96] and merge the edited face regions with input upper-body images, this often creates artifacts. In contrast, our method generates changes in the visible body and hands that are consistent with the input audio. See more examples in Supp. Video.

Figure 5. **Showcasing model diversity**. VLOGGER is stochastic and can generate multiple videos for the same input image. We condition generation on a text prompt, with the input image shown top-left. Row a) shows increasing diversity in the temporal domain (left-to-right). Notice that the model has low deviation (blue) in the background pixels and has significant variation (green and red) in blinking, facial expressions, head pose and even upper-body areas. Row b) shows the same timestep for 4 different videos generated, all of good visual quality.

were available. In Fig. 6, we show that by fine-tuning our diffusion model with more data, on a monocular video of a subject, VLOGGER can learn to capture the identity better, *e.g.* when the reference image displays the eyes as closed.

Figure 6. Qualitative results on model personalization. Finetuning our model [63] on a single video of a user supports more veridical synthesis over a wide range of expressions.

5. Conclusion

We have presented VLOGGER, a methodology for human video synthesis, including both face and body, from a single input image, conditioned by audio or text. VLOGGER is built as a temporal extension of control-based diffusion models, with underlying scaffolding based on 3d human head and body pose representations, which generates high quality animations of variable length. We introduce a diverse and large scale dataset (one order of magnitude larger than previous ones), and validate the performance of our model on this and multiple other repositories, showing that VLOGGER outperforms previous state-of-the-art methods

on the task of talking face generation, and that our approach is more robust on different diversity axes.

SUPPLEMENTARY MATERIAL

In this supplementary material, we first provide background for the head model SPHEAR [6], being used to represent expressions and head pose, and an overview of diffusion models. We then provide more details on the processing and statistics of the MENTOR dataset, and extensive implementation details. We finally include more results and a supplementary video summarizing our contributions and results.

A. Background

In this section we provide more background on the statistical head model SPHEAR [6] and diffusion models [31]. Please refer to the original papers for all details.

A.1. SPHEAR Head Model

We use a recently introduced statistical body model called SPHEAR [6] to represent the pose and the shape of the human head. The model has been trained end-to-end in a deep learning framework, using a large corpus of human head shapes and facial expressions $(\boldsymbol{\beta}_s, \theta^e)$, which are represented using deep variational auto-encoders. The mesh consists of $N_v = 12,201$ vertices and $N_f = 24,318$ faces. The model is rigged and has a number of $J = 15$ joints including a root joint. The other degrees of freedom include the neck, head, tongue, left and right eyeballs, lower and upper eyelids. It uses pose encodings represented as 6d rotations [105]. SPHEAR relies on linear blend-skinning, pose-space deformations and facial expression correctives for mesh generation from latent codes [6, 87].

A.2. Diffusion Models

VLOGGER is based on two diffusion models to first synthesize realistic human motion and, based on this, generate realistic videos of the person moving. We therefore give a brief explanation of diffusion models in our setting. Let $\mathbf{x}^0 \sim q(\mathbf{x}^0)$ denote a data point from a distribution q. In order to learn $p_\theta(\mathbf{x}^0)$ which can model $q(\mathbf{x}^0)$, diffusion probabilistic models consider a forward and a reversed process.

The forward process gradually deconstructs $\mathbf{x}^0$ by injecting Gaussian noise in each step $t \in T$ such that

$$q(\mathbf{x}^t|\mathbf{x}^{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}^{t-1}, \beta_t\mathbf{I}), \tag{3}$$

for a given number of steps T, hyperparameters β and the diagonal matrix $\mathbf{I}$. This is equivalent to a closed-form solution that can be sampled directly for step t in $\mathbf{x}^t = \sqrt{\alpha_t}\mathbf{x}^0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\hat{\alpha}_t = 1 - \beta_t$ and $\alpha_k = \prod_{i=1}^{t} \hat{\alpha}_i$.

During the reversed diffusion process, a model $p_\theta(\cdot)$ with parameters θ is trained to approximate the inverse process and generate samples from Gaussian Noise $\mathbf{x}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In our framework, the video generation network follows [64] by predicting ϵ_t as formulated in Ho et al. [31]. However, the motion generation network predicts the signal itself, *e.g.* $p_\theta(\mathbf{x}^t, t) \mapsto \mathbf{x}^0$, since it was very difficult to converge for such model with the original formulation.

B. MENTOR dataset

The MENTOR Dataset consists of a large number videos that are available online under permissive licenses, containing a single speaker, and for a duration of 10 seconds. This is equivalent to 240 frames at 24 fps, with an audio sample rate of 16 kHz. The videos contain a large number of subjects and feature unique diversity in terms of perceived attributes such as skin tone, age, gender. Moreover, the videos have different visibility conditions, including tight faces (as in most currently available datasets), upper-body or even hands. Next, we provide more details about the fitting pipeline and discuss the statistics of MENTOR with respect to currently available benchmarks.

B.1. Preprocessing

In this section, we extend on the details provided in the main paper and provide more information about the preprocessing step. For each video, we first obtain the pixel difference in consecutive frames and remove videos whose difference has a standard deviation smaller than 3 (person barely moving) or larger than 10 (background and foreground moving drastically). We next estimate the 2D landmarks and remove videos where the facial and body landmarks differ more than 1 pix in consecutive frames, on average. Finally, we also remove videos where the face has not been detected in at least one frame or hand detection is irregular, which most often happens when people are manipulating objects.

We then fit the statistical head model SPHEAR [6] in each video to represent head and face motion expressions, proposing a two-stage pipeline to obtain expressive yet temporally consistent fits. We first rely on an in-house facial landmark detector that provides a sparse semantic signal per-frame. Next, we use a trained landmark lifter that goes from 2d landmarks to initial per-frame head parameters (similar to [22, 23]). We then finetune the previous frame predictions by using common face shape parameters, and minimizing a landmark projection error, a temporal consistency loss (difference in SPHERE parameters between consecutive frames) and an expression prior [87]. We weigh these losses with $1e^0$, $1e^0$ and $1e^{-1}$ respectively. This process follows a non-linear quasi-Newton optimization (L-BFGS) framework for 3000 steps on CPU, which takes approximately 300 seconds for each 240-frame video. We adopt a full-perspective camera, with the initial diago-

	Properties						Size		Body visibility			
Source	Has audio	Subj. Diversity	High-Res	FPS	SR	In-the-Wild	# Subjects	# Hours	Face	Body	Hands	
Lab	✓	✗	✓	30	48 kHz	✗	24	7	✓	✗	✗	RAVDESS [44]
TED	✓	✓	✗	25	16 kHz	✓	1k	165	✓	✗	✗	LRW [11]
TED	✓	✓	✗	25	16 kHz	✓	5k	400	✓	✗	✗	LRS3 [2]
Lab	✓	✗	✓	30	48 kHz	✗	60	15.8	✓	✗	✗	MEAD [78]
BBC	✓	✓	✗	25	16 kHz	✓	6.1k	2400	✓	✗	✗	VoxCeleb2 [14]
YouTube	✗	✓	✓	25	16 kHz	✓	222	1.5	✓	✗	✗	TalkingHead-1KH [81]
YouTube	✓	✗	✓	25	16 kHz	✓	300	40	✓	✗	✗	HDTF [99]
YouTube	✓	✓	✓	24	16 kHz	✓	800k	2200	✓	✓	✓	**MENTOR**

Table 5. **Statistics of MENTOR in comparison to currently available datasets.** While most datasets are collected from the same sources (*e.g.* TED or YouTube), there is no guarantee this leads to a better diversity of subject distribution in terms of perceived age, gender or skin tone. We assume that datasets with more than 1k subjects will be moderately diverse, however, in the case of LRW and VoxCeleb2, might limited to the subject distribution appearing in TED and BBC. Some datasets have focused on collecting talking humans with emotions labels [44, 78] and require actors, which limits the amount of videos and subjects, and restricts the dataset to lab recordings. When considering size, the number of different identities in MENTOR is **two orders of magnitude larger** than the closest competitor and three orders of magnitude when considering high-resolution benchmarks. Moreover, communication is much more than talking faces, and MENTOR is the first dataset of talking humans that considers faces, upper-body and hands, going a step further than previous benchmarks. See Fig. 7 or the Supplementary Video for examples of videos in the dataset.

nal field-of-view being estimated from the input video and finetuned further during the previous optimization. Fig. 7 shows an example of images in the MENTOR dataset with head fits and predicted 2D landmarks. This process resulted in a large scale training set with unique properties that we discuss more in detail in the next section.

B.2. Statistics

We report statistics and properties of currently available datasets in comparison to MENTOR in Tab. 5. MENTOR goes a step further than previous benchmarks, being the first to consider talking humans with upper-body and hands. In this work we argue that these are important when humans communicate, and have not been well represented in available datasets, thus hope that our work and dataset will enable researchers to move in the direction of full-body moving and talking avatars. We also report other properties of all datasets, such as subject diversity, even though this will be limited depending on the video source and their original subject distribution. Finally, the MENTOR dataset contains more than 2200 hours of high-resolution videos in-the-wild, with a number of subjects two orders of magnitude larger than the closest competitor [14], whose videos are at lower resolution, and three orders of magnitude when considering high-resolution benchmarks [81, 99].

C. Implementation Details

C.1. Audio-Driven Motion Generation

Our audio-to-motion network is implemented following a transformer decoder [73] architecture that takes the Mel-Spectrogram obtained at 256 mel bins with a Hann window of length 250 ms. The model is formed of four transformer layers, each with four heads and key, value and query size of 256. We use a two-layer MLP after each attention block with GELU activations [28] and LayerNorm [3]. After the final block, we have different fully-connected layers for face parameter estimation, body and hands. The diffusion timestep t is standardly embedded as a 128-vector via a positional encoding. We follow standard DDPM [31] for 1000 steps and a cosine noise schedule. At training time, we mask the audio conditioning with 0.2 probability as in [30] to follow a classifier-free guidance framework at test time. In this regard, we run inference with a guidance weight of 5, which leads to better LSE results as shown in the Ablation Study of the main document. Our dataloader samples random videos with a variable length between 3 and 10 seconds, and uses a causal mask that enforces it to attend only to previous time steps. The model was trained at batch size 512, learning rate $1e-4$ with decay rate of 0.99 every 4000 steps, for 600k steps with loss weights $\mathcal{L}_{\text{diff}} = 1$ and $\mathcal{L}_{\text{temp}} = 42$ for the diffusion and temporal losses respectively.

C.2. Video Generation

Our generative model of videos extends a version of Imagen [64] trained on internal data, which generates base images at 128×128 resolution. We make a zero-initialized trainable copy of the encoding blocks, and introduce a new temporal convolution layer with kernel size 3, after the first Swish Activation (See [64] for details on the original architecture) in each downsampling block. For the super-resolution modules, we extend the efficient U-Net architecture. We train the models at $\texttt{lr} = 5e{-}5$, for $280k$ steps with $\texttt{bsize} = 128$, and use a cosine noise schedule as in [53].

The network takes the reference image, with reference head pose/shape information, and the new head poses as controls. Therefore, we further preprocess the videos in the MENTOR dataset by selecting random 8-frame clips, for which we select a random reference frame that is at

Figure 7. **Samples of the MENTOR Dataset.** In this figure we show six videos of the MENTOR Dataset that summarize the distribution of humans in the videos. The videos are in-the-wild and show very diverse backgrounds and subjects, expressions, cloth styles, lighting, motion blur and body visibility (*e.g.* face, upper-body or hands). The videos whose body is not accurately estimated are automatically removed.

least three seconds apart from the beginning or end of the groundtruth frames. This ensures that the head pose and expression are more diverse and trains the model to generalize better in case of challenging reenactment examples.

D. Additional Results

We provide more qualitative results in Figures 8 and 9, which include the input image on the left for a subject in each row, and the generated sequence on the right. While the initial head pose is similar to the one in the input image, the model stochastically generates realistic head motion in a temporally consistent manner, leading to a set of varying set of images with more apparent differences at the end of the video (right-most image). The two figures show significant diversity in terms of subjects, body visibility and input pose of the head and body, validating the capacity of VLOGGER in images in-the-wild and without requiring no pre-processing (*e.g.* detection of cropping faces).

We also summarize the contributions of our paper in the supplementary video, with several examples and comparisons against the most recent State-of-the-Art methods [46, 96]. In the video, we show intermediate examples of motion-to-video generation to prove that our video generation model can tackle challenging motion of faces and even body or hands. Furthermore, we extend this approach in the case of reenactment, by transferring the head motion from a driving video to each target person. Our approach in this case is to transfer the expression and pose parameters from the head model [6] to the new video. We also discuss new examples of generated videos given the same person and audio, and current limitations of our work.

References

[1] Martin Adam, Michael Wessel, and Alexander Benlian. Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 31(2):427–445, 2021. 2

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 10

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 10

[4] Bard. Bard: A large language model from google ai, 2023. 2

[5] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 3

[6] Eduard Gabriel Bazavan, Andrei Zanfir, Teodor Alexandru Szente, Mihai Zanfir, Thiemo Alldieck, and Cristian Sminchisescu. Sphear: Spherical head registration for complete statistical 3d modeling. *3DV*, 2024. 3, 4, 5, 6, 9, 12

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 3

[8] Stella Bounareli, Vasileios Argyriou, and Georgios Tzimiropoulos. Finding directions in gan's latent space for neural face reenactment. *BMVC*, 2022. 3

[9] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces. In *ICCV*, pages 7149–7159, 2023. 2, 3

[10] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, pages 7832–7841, 2019. 3

[11] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016. 10

[12] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *ACCV-W*, 2016. 6

[13] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that?, 2017. 3

[14] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 10

[15] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *CVPR*, 2019. 2, 3

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 6

[17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 3

[18] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *ICCV*, pages 14398–14407, 2021. 3

[19] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *CVPR*, pages 18770–18780, 2022. 2, 3

[20] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *CVPR*, pages 5609–5619, 2023. 2

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[22] Ivan Grishchenko, Valentin Bazarevsky, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, Richard Yee, Karthik Raveendran, Matsvei Zhdanovich, Matthias Grundmann, and Cristian Sminchisescu. Blazepose ghum holistic: Real-time 3d human landmarks and pose estimation. *arXiv preprint arXiv:2206.11678*, 2022. 4, 5, 9

[23] Ivan Grishchenko, Geng Yan, Eduard Gabriel Bazavan, Andrei Zanfir, Nikolai Chinaev, Karthik Raveendran, Matthias Grundmann, and Cristian Sminchisescu. Blendshapes ghum: Real-time monocular facial blendshape prediction. *arXiv preprint arXiv:2309.05782*, 2023. 5, 9

Figure 8. **Additional qualitative results from our method**. Given the input images on the left, VLOGGER generates the following sequences of the person speaking naturally.

[24] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *CVPR*, pages 1505–1515, 2023. 2

[25] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, pages 5784–5794, 2021. 3

[26] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, pages 10893–10900, 2020. 3

[27] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 3

[28] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 10

[29] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 5, 6, 7

[30] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 10

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 9, 10

[32] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[33] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 5

[34] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking

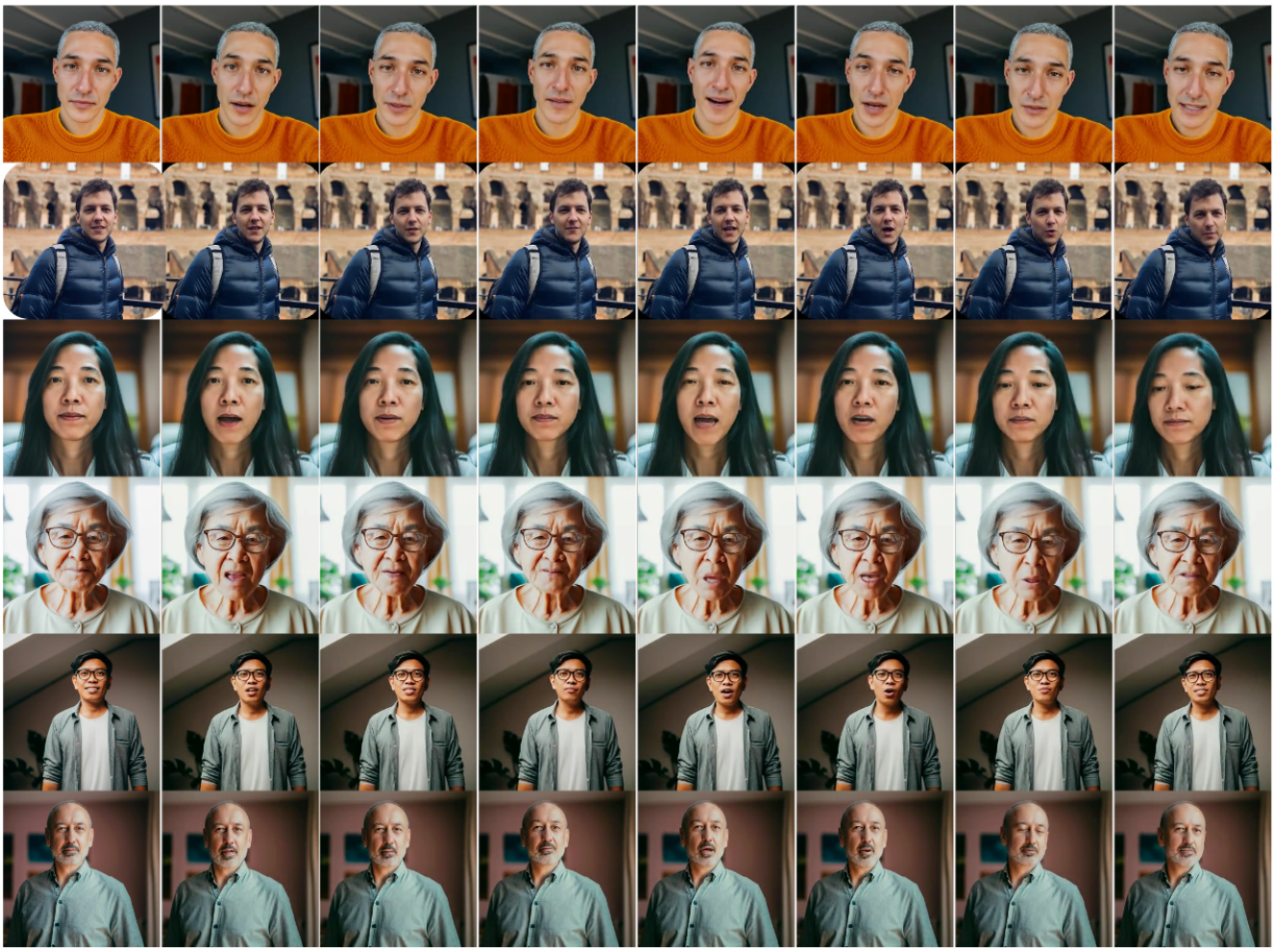


Figure 9. **Additional qualitative results from our method**. Given the input images on the left, VLOGGER generates the following sequences of the person speaking naturally.

head video generation. In *CVPR*, pages 3397–3406, 2022. 3

[35] Gee-Sern Hsu, Chun-Hung Tsai, and Hung-Yi Wu. Dual-generator face reenactment. In *CVPR*, pages 642–650, 2022. 2, 3

[36] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *IJCV*, 127:1767–1779, 2019. 3

[37] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *SIGGRAPH*, 2022. 3

[38] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 3

[39] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 3, 5

[40] Christos Kyrlitsias and Despina Michael-Grigoriou. Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality*, 2:786665, 2022. 2

[41] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *TOG*, 36(6):194–1, 2017. 3

[42] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *ECCV*, pages 106–125. Springer, 2022. 3

[43] Yunfei Liu, Lijian Lin, Fei Yu, Changyin Zhou, and Yu Li. Moda: Mapping-once audio-driven portrait animation with dual attentions. In *ICCV*, 2023. 2

[44] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal

expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 10

[45] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live Speech Portraits: Real-time photorealistic talking-head animation. *ACM Transactions on Graphics*, 40(6), 2021. 3

[46] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. *arXiv preprint arXiv:2301.01081*, 2023. 2, 6, 7, 8, 12

[47] James Manyika. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2023. 2

[48] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[49] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *Signal Processing Letters*, 20(3):209–212, 2012. 6

[50] Sara Moussawi, Marios Koufaris, and Raquel Benbunan-Fich. How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. *Electronic Markets*, 31:343–364, 2021. 2

[51] Niranjan D Narvekar and Lina J Karam. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *International Workshop on Quality of Multimedia Experience*, pages 87–91. IEEE, 2009. 6

[52] Niranjan D Narvekar and Lina J Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *Image Processing*, 20(9):2678–2683, 2011. 6

[53] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 10

[54] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, pages 7184–7193, 2019. 2, 3

[55] OpenAI. Gpt-4 technical report, 2023. 2

[56] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee, 2009. 3

[57] Gabriele Pizzi, Virginia Vannucci, Valentina Mazzoli, and Raffaele Donvito. I, chatbot! the impact of anthropomorphism and gaze direction on willingness to disclose personal information and behavioral intentions. *Psychology & Marketing*, 40(7):1372–1387, 2023. 2

[58] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM International Conference on Multimedia*, pages 484–492, 2020. 2, 3

[59] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: One-shot anatomically consistent facial animation. In *IJCV*, 2019. 3

[60] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, pages 13759–13768, 2021. 3

[61] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *ICCV*, pages 1173–1182, 2021. 2, 3

[62] Eileen Roesler, Dietrich Manzey, and Linda Onnasch. A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, 6(58): eabj5425, 2021. 2

[63] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arxiv:2208.12242*, 2022. 7, 8

[64] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022. 4, 5, 9, 10

[65] Anna-Maria Seeger, Jella Pfeiffer, and Armin Heinzl. Texting with humanlike conversational agents: Designing for anthropomorphism. *Journal of the Association for Information Systems*, 22(4):8, 2021. 2

[66] Lennart Seitz, Sigrid Bekmeier-Feuerhahn, and Krutika Gohil. Can we trust a chatbot like a physician? a qualitative study on understanding the emergence of trust toward diagnostic chatbots. *International Journal of Human-Computer Studies*, 165:102848, 2022. 2

[67] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. *arXiv preprint arXiv:2309.11306*, 2023. 2, 3

[68] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. *arXiv preprint arXiv:2301.03396*, 2023. 3

[69] Wang Suzhen, Li Lincheng, Ding Yu, Fan Changjie, and Yu Xin. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *IJCAI*, 2021. 3

[70] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *ICCV*, pages 20621–20631, 2023. 2, 3

[71] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016. 2, 3

[72] ttscloud. Text-to-speech - google cloud. https://cloud.google.com/text-to-speech, 2019. 3

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3, 10

[74] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi

Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 3, 5

[75] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *IJCV*, 128:1398–1413, 2020. 3

[76] Mattias Wahde and Marco Virgolin. Conversational agents: Theory and applications. In *HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation*, pages 497–544. World Scientific, 2022. 2

[77] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *CVPR*, pages 14653–14662, 2023. 2

[78] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 10

[79] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 6

[80] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *AAAI*, pages 2531–2539, 2022. 6, 8

[81] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 2, 4, 6, 7, 10

[82] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, pages 10039–10049, 2021. 3

[83] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, pages 670–686, 2018. 3

[84] Xiuzhe Wu, Pengfei Hu, Yang Wu, Xiaoyang Lyu, Yan-Pei Cao, Ying Shan, Wenming Yang, Zhongqian Sun, and Xiaojuan Qi. Speech2lip: High-fidelity speech to lip generation by learning from a short video. In *ICCV*, pages 22168–22177, 2023. 2

[85] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023. 3, 5

[86] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *CVPR*, pages 12780–12790, 2023. 2, 3

[87] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. *CVPR*, 2020. 3, 9

[88] Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. Face2face ρ: Real-time high-resolution one-shot face reenactment. In *ECCV*, pages 55–71. Springer, 2022. 2, 3

[89] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 3

[90] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, JinZheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023. 3

[91] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *CVPR*, pages 469–480, 2023. 2, 3

[92] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40 (4):1–13, 2021. 6

[93] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *ECCV*, pages 85–101. Springer, 2022. 3

[94] Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. Talking head generation with probabilistic audio-to-visual diffusion priors. In *ICCV*, pages 7645–7655, 2023. 2, 3

[95] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 4

[96] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, pages 8652–8661, 2023. 2, 6, 8, 12

[97] Yunxuan Zhang, Siwei Zhang, Yue He, Cheng Li, Chen Change Loy, and Ziwei Liu. One-shot face reenactment. *arXiv preprint arXiv:1908.03251*, 2019. 2, 3

[98] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, pages 3661–3670, 2021. 3

[99] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, pages 3661–3670, 2021. 2, 3, 6, 7, 10

[100] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, pages 3657–3666, 2022. 3

[101] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023. 2

[102] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, pages 9299–9306, 2019. 3

[103] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, pages 4176–4186, 2021. 3

[104] Qi Zhou, Bin Li, Lei Han, and Min Jou. Talking to a bot or a wall? how chatbots vs. human agents affect anticipated communication quality. *Computers in Human Behavior*, 143:107674, 2023. 2

[105] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 9

[106] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020. 6